图书情报工作
第63卷 第7期 2019年4月

基于知识元的中文文本层级分割*

王忠义[1] 沈雪莹[1] 黄京[2]
[1] 华中师范大学信息管理学院 武汉 430079 [2]武汉职业技术学院 武汉 430074

摘要：[目的/意义]为帮助用户检索到完整的、粒度大小适当的知识单元,满足用户多粒度的知识需求。[方法/过程]提出一种基于知识元的文本层级分割方法。该方法首先对知识元的类型及其描述规则进行分析;然后依据知识元描述规则识别实体资源中的各类型知识元,并将所有的知识元和知识元之间的衔接句视为一个类;最后基于fisher分割算法对该类进行逐级二分,直到识别出所有的主题为止,确定分割边界,实现文本层级分割。[结果/结论]基于知识元的中文文本层级分割方法,一方面使得文本分割单元从句子扩展为知识元,提高分割时的效率,另一方面将知识服务的控制单元从文献深入到以知识元、知识元集合为单位的知识块,按需为用户提供相关知识服务,使数据检索、信息检索向知识检索迈进,提高知识获取效率,实现信息服务向知识服务的转型。

关键词：知识元识别 聚类 层级分割
分类号：G254
DOI:10.13266/j.issn.0252-3116.2019.07.013

1 引言

随着网络技术的发展,网络信息资源快速膨胀,人们工作、生活节奏的加快,以往的检索系统很难较好地满足用户的知识需求,这是由于传统的检索系统通常以篇章为单位进行检索,其返回结果通常是整篇文献,导致信息过载,使得用户不得不花费大量的时间和精力逐篇阅读文献才能查找定位到文献中所蕴含的知识,浪费用户大量的时间。事实上,用户检索的主要目的是能够及时查找到满足其知识需求的粒度大小适当的知识块,为解决这一问题,实现深入文献内部的多粒度知识服务,这就需要对文本进行多粒度的层级分割。为此,本文提出了基于知识元对文献进行层级分割的方法,帮助实现基于知识元、知识元集合的多粒度知识服务,使得用户在进行知识检索时,能够准确得到其所需要的知识模块,而不是整篇文档,使数据检索、信息检索向知识检索迈进,提高知识获取效率,实现信息服务向知识服务转型。

2 文本分割研究现状

文本分割是指在一个书面文档中自动识别具有独立意义的知识单元之间的边界的过程,其在信息检索和文本智能处理等领域有着重要的应用[1]。目前,国内外学者对文本分割的相关研究已经有了初步的成果。一般而言,文本分割大致分为两大类,一类是线性分割,即将文本分成连续片段,不考虑文本内在结构,第二类是层级分割,即将文档迭代分割为更精细的具有层次结构的片段。

2.1 文本线性分割研究现状

目前,常用的线性分割方法主要分为以下几类:基于语言特征的分割、基于词汇集聚的分割、基于主题模型的分割以及使用以上不少于两种的混合方法的分割。

基于语言特征的分割方法是从文本中提取词汇特征,研究它们与主题片段首尾之间的关系,进而来确定主题边界。J. C. Reynar提出了一种基于词汇特征的文本分割算法,即单独或组合使用特征来识别若干文档中的主题转换[2]。邹箭和钟茂生等考虑到中文文本的特殊性,提出了针对中文的文本分割模型,根据语料库和词典对词语的相关度进行计算来分析句子之间的相关度,并进行分割[3]。然而,这种方法只适用于一些含有明显的形式化信息的特定文本,无法适用各种文

* 本文系教育部人文社会科学研究青年基金"数字图书馆馆藏资源多粒度层级主题分割研究"(项目编号:16YJC870003)研究成果之一。
作者简介:王忠义(ORCID:0000-0001-8945-783X),副教授,博士,硕士生导师,E-mail:wzywzy13579@163.com;沈雪莹(ORCID:0000-0002-2944-4399),硕士研究生;黄京(ORCID:0000-0003-2938-8507),副教授。
收稿日期:2018-06-22 修回日期:2018-10-11 本文起止页码:105-115 本文责任编辑:杜杏叶

本，移植性较差。

基于词汇集聚进行分割的思想来源于 M. A. K. Halliday 和 R. Hasan，他们将词汇聚集的表现归纳为词的重复或变相重复以及词汇之间的语义联系[4]。H. Kozima 对于如何测量词汇紧凑度提出了一种词汇集聚图（LCP）的文本线性分割方法[5]。J. C. Reynar 和 M. A. Hearst 基于该理念分别提出了 Dotplotting 算法与 TextTiling 算法，Dotplotting 算法主要完全依赖于单词重复来找到紧密的主题相似区域，进而识别边界点[6]，而 TextTiling 算法主要基于单词重复和单词矢量的相似性计算文本单元之间的相似性，来确定边界[7]。F. Y. Y. Choi 基于 Dotplotting 算法提出了 C99 算法，该算法是建立在文档中所有句子的相似矩阵，即通过计算文本中句子之间的余弦相似度构建相似度矩阵，对相似矩阵进行排序进而优化，然后使分割单位的内部密度最大化，进而实现分割[8]。J. M. Ponte 和 W. B. Croft 利用词汇之间的语义联系，提出了一种基于局部上下文分析的文本分割方法，用来查找与每个句子相关的单词和短语[9]。其他基于词汇聚集的方法，较为明显的就是基于词汇链的文本分割，J. Morris 认为词汇链的首尾与文本结构具有对应关系，可以计算词汇链，以此度量片段边界[10]。上述分割方法完全基于文本中所包含的信息进行分割。但是，当特定主题中的句子由于使用同义词而不共享通用词并允许语义上相关的词表示主题连续性时，上述分割算法可能无法确定可靠的边界。

为了克服上述分割中的不重复问题，主题模型也受到关注，其不仅通过语义信息来进行文本分割，而且还用于减少单词向量的稀疏性。F. Y. Y. Choi 等人提出了一种通过潜在语义分析（LSA）来估计句子间相似性的线性分割的方法[11]。石晶提出了基于概率潜在语义分析（PLSA）模型和基于潜在狄利克雷（LDA）模型的文本分割方法，并进行比较，通过实验发现基于 LDA 模型进行分割的准确度比 PLSA 高[12-13]。M. Riedl 和 C. Biemann 介绍了将 LDA（潜在狄利克雷分配）主题模型并入文本分割算法的一般方法，结果表明，主题模型添加的语义信息显著提高了 TextTiling 和 C99 两种基于词的算法的性能[14]。J. Eisenstein 和 R. Barzilay 提出一种新的贝叶斯方法来进行无监督的主题分割，即通过将每个主题段中的单词用与该段相关的多项语言模型绘制，并进行建模，可以将词汇内聚置于贝叶斯概率模型中，在此模型中观察最大可能产生一个词汇内聚的分割[15]。P. Mulbregt 等人引入用于文本分割的隐马尔可夫模型来进行主题检测和跟踪[16]。尽管主题模型能够很好地提高文本分割的性能，但是，此类方法的主题个数则普遍依赖于人工，针对不同的数据集，最优主题个数不同。

很多学者在研究文本分割时，会将两种或更多的方法组合起来以获取更好的分割效果。T. Brants 和 F. Chen 等人提出将概率潜在语义分析（PLSA）模型与相邻块之间的相似度值选择分割点的方法相结合来选择[17]。M. Riedl 等人提出了将 TextTiling 和 LDA 模型相结合的 TopicTiling 方法，即通过 LDA 模型获得最终的主题分布，使主题模型对文本的表示更加稳定[18]。M. Y. Kan 提出了将词汇衔接的特征与布局识别中的元素进行整合，以建立一个复合框架，进而使用框架来计算文档结构的方法[19]。

2.2 文本层级分割研究现状

尽管人们普遍认为大多数文档是具有层次结构的，但是有关研究文本层次分割的文献相对较少。Y. Yaari 提出了一种无监督的分层主题分割方法，如同在 TextTiling 中，使用余弦相似度测量内聚力，并使用聚类来形成段落上的树状图，然后使用启发式算法将树形图转换为分层分割[20]，但是这种启发式方法通常是易碎的，因为其包含的许多参数需要手动调节。为了克服这些问题，J. Eisenstein 提出了一种新型无监督的方法来执行文本分层分割，该方法集成了贝叶斯概率框架，利用多尺度凝聚力来实现分层分割[21]。然而，由于段落层次的词汇数量稀少，该方法并没有扩展到更细粒度的片段，比如段落，这样将有必要明确语篇连接词和词汇语义描述。Y. W. Teh 等基于层次分割提出了分层狄利克雷过程（HDP）模型[22]。李天彩和王波等提出了一种基于 HDP 模型运用 C99 分割算法进行文本层次分割，即首先使用 HDP 模型对文本进行向量空间表示，然后将主题向量用于 C99 分割算法来实现分割[23]。该方法有助于优化文本分割，但是对于较短段落的分割错误率较高。

综上所述，学术界对文本分割的研究不断改进，但是对文档进行层级分割的研究相对较少，然而数字图书馆的数字馆藏资源大都体现出层级结构，为实现对数字馆藏资源的多粒度层级组织，需要在此基础上对文档层级分割方法展开进一步深入研究。此外，现有的文档分割方法通常以一句话、几句话或一个段落作为最小分割单元，这些分割单元无法保障自身在逻辑上是一个完整的知识单元，要么粒度过细割裂了知识之间的内在联系，要么粒度过粗模糊了知识之间的界

限，这些现象都将导致文档分割的错误率较高。而知识元作为具有相对独立意义的不可再分的最小知识单元，是构成知识结构的基元，因此，以知识元为单位进行文档层级分割将可以有效解决上述问题。基于上述分析，本文在前人研究的基础上，提出基于知识元的中文文本层级分割方法，以知识元为基元来对文档进行层级分割，这将有利于将数字图书馆知识组织的单位由粗粒度的文献单元深入到细粒度的知识元层次。

3 知识元的识别

3.1 知识元的描述规则

为实现基于知识元的文档层级分割，首先要识别文档中的知识元。本文基于规则的方法对文本中的知识元进行识别。由于不同类型的知识元描述规则不同，为了统计规则的完整性，需要分析知识的类型。目前针对知识元的分类，不同学者有着不同的见解。温有奎认为知识元的类型主要分成两大类，即描述型和过程型，前者包括信息型、名词解释型、数值型、问题描述型和引证型，后者包括步骤型、方法型、定义型、原理型和经验型[24]。张静根据对中小学各学科中所含知识的研究，将知识元分为：概念型、原理型、方法型、事实型和陈述型[25]。原小玲根据知识元表达的内容将知识元分为理论与方法型、事实型和数值型[26]。赵蓉英将知识元划分为陈述型和程序型，前者包括事实知识元、定义知识元和结论知识元等陈述型内容，后者包含方法知识元和关系知识元等具有内在结构的内容[27]。

综上所述，不难发现有些学者对知识元的分类过细，不同类型知识元之间存在交叉，如温有奎提出的名词解释型和定义型，这两个句型结构以及描述规则大致是一样的。有些学者对知识元分类过粗，未能包含所有的知识元类型。因此，本文结合前人的观点，基于文献的内容表达方式将知识元分为描述型和过程型，前者包括概念知识元、事实知识元和数值知识元，后者包括方法知识元和关系知识元。知识元的分类及其关系如图1所示：



图1 知识元的分类及相关关系

本文选取期刊论文为研究对象，分别选择13门学科中排名前5的核心期刊，并在每门学科的每个核心期刊中选取近五年被引量前10的文章，即共计650篇文献作为训练语料。然后，对这些期刊论文进行文档解析，转换成纯文本，分别提取出中文摘要、关键词以及正文部分。接着从训练语料的关键词信息和摘要信息中提取描述文献主题的关键词集合，作为初始的术语表。而后根据得到的术语表，对语料中包含术语的知识元语句进行抽取，并过滤掉知识元语句中的领域词，得到句子的线性句式结构。最后，人工审核、校对，并依据知识元类型对句式结构进行归类汇总，生成各知识元类型的描述规则。其流程如图2所示：

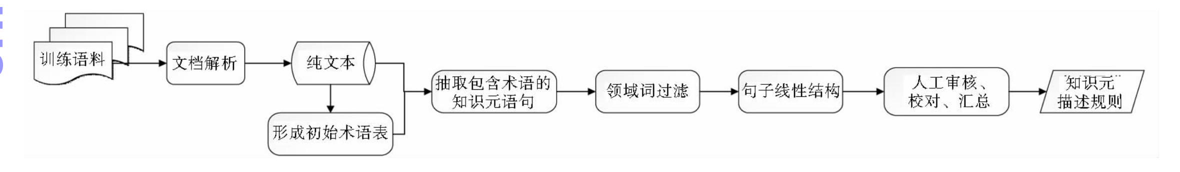


图2 生成知识元描述规则流程图

3.1.1 描述型知识元的描述规则

（1）概念知识元描述规则。概念知识元是一种较为抽象概括的、有组织的知识性描述，也是对某个对象的本质特征或外延的简要说明，即表明某一学科领域的对象是如何加以组织的，如何体现出系统一致的方式以及名词术语等的解释、特点和作用。其描述中含有的特征词一般有“是”“是指”和“定义为”等，有关该类知识元的描述规则总结见表1。

（2）事实知识元描述规则。事实知识元包括自然、社会存在和演变的事实信息，一般是对研究的背景、现存问题以及专家或者国际观点、认识等信息的描述。本文参照文献[27]将事实知识元分为观点型、序列型、直述型、分析预测型和事件型。观点型事实知识元一般是对事实发表观点的结构的描述，句式简单，其句式结构一般为：观点发出者＋观点内容＋观点释义。序列型事实知识元是对事实进行形式化的描述，由序列性的连接词构成，条例清晰，结构分明。其特征词一般为数字连接词和表达序列关系的连接词。直述型即直接叙述型，对事实进行直接的陈述性表达，不存在句式上的修辞，一般没有特定的描述规则。其句式结构一般为：主语＋论述内容，其中关键词术语表内容作为主语，后衔接论述内容。预测型事实知识元是根据事

表1　概念知识元描述规则

概念知识元描述规则
Term + 是指/是/指/称/称作/称为/称之为/叫/叫做/内涵是/定义是/定义为/本质即/实质是/有 N 层含义为/概念/意思/定义如下/即/又称…
所谓/关于 + Term + [是/指/是指/即]…
…被称为/被定义为 + Term
Term + 具有以下特点/优点/缺点/不足/缺陷/好处/特征:…
Term + 具有/尚存/存在/融合了/综合了/兼顾了/克服了/避免了... 的特点/特征/特性/问题/优点/缺点/不足/问题/局限/缺陷
Term + 克服/避免/无法/不能/适用于…
Term + 对…来说是…挑战
…表明/与…相比,… + Term + (在…方面) + 更具/好处是…
Term + 可以/能够/能/不能/有助于/有利于/有益于/提高了/打破了/减少了/常用来/主要用于/侧重于/揭示/导致/影响/带来/改变了/阻碍/限制/帮助/用于/适用于/可用于/被应用到/被用于…
Term + 不仅要/不仅能…还(并且)要/还(并且)能…
Term + 的职能/作用/功能/(主要)用于/用来/用作/用做…
Term + …在…有着/得到…的应用/功能/效用/意义
Term + 极易 + 带来/造成/提高…(影响/效率/作用)
Term + (具有 N 个显著的特征/细分为 N 个层次/主要有/包含/包括/由 N 部分构成(组成))…,其中,第一/一/一是/(1)/①/a…;第二/二/二是/(2)/②/b…;第三/三/三是/(3)/③/c…

实进行的推论,其特征词主要是"发现""根据"等相关与预测性相关的词。事件型事实知识元是对一个事件的完整性陈述,一般涉及事件发生的地点、时间、人物等。其句式结构一般为:时间点 + 主语 + 地点 + 事件,主语 + 时间点 + 地点 + 事件。有关该类知识元的描述规则如表2所示:

表2　事实知识元描述规则

事实知识元描述规则
主语 + (在)…中(时)指出/阐述/提出/提到/表示/表达/认为/说/明确/证明/发现/强调/探讨...(是/存在/有)…
[从…角度]/[以…为例]/[利用…的方式]/[以…为理论起点] + 提出/指出…
基于/关于/针对/对于…提出/认为/声明/称/表示/指出…
主语 + (对)研究了…,在…(基础上)提出…(改进)
实验证明了/实验结果表明 + …(Term)…
… + (主语,时间)
当前,…现实/事实/现象 + 就是…
(从/由/分析/通过/根据/依据/综合…(角度)),看出/了解到/得出/可知/反映/预计/出现/发现/说明/可知/可见…
基于(此)…分析(发现)…
未来…的(发展)趋势 + 是 + …
今后…,预计会…
(…年…月…日)/世纪/年代/朝代,于 + 地点…发生/出席有/创造/遭遇/诞生/召开…
(…年…月…日)/世纪/年代/朝代,有关… + 声明/表明/明确…
主语 + (…年…月…日)/世纪/年代/朝代/近几年 + 推出/声明/上线/创造/确定/提供/印发/发布/宣布/出台…

(3)数值知识元描述规则。数值知识元是从数字的角度来阐述某事物或事件的性质及其运动规律的认识,如用长度、高度、货币、时间、重量、百分比等以数值形式存在的完整的描述。一般来说,文献中描述数值信息的一般分为三大类,即基数类数值信息、数量类数值信息以及数值知识元,其中,第二类是由第一类数值信息的基础上加上量词或者符号组成的,第三类是在第二类的基础上加上句子中其他的描述成分所组成。从上述分类中可以看出,数值知识元是在数量类数值信息的基础上形成的,与数值信息直接相关的词不多,运用数值来进行描述说明的相关词主要就是数字、量词和特定符号,即数值 + 单位。描述数值的句子中,时间表述一般为:(…年…月…日至…年…月…日)/(…年…月…日)/(截至/截止/日期/时间为)/年代/朝代/世纪;单位一般为:个/篇/件/元/条/名/位等;指标表述一般为:论文/文献/文档/专利/结果。本文可以借鉴温有奎的数值知识元的抽取方法,首先识别出包含数值的句群,判断识别的句群中的数值是否具有数值价值,然后人工对数值知识元的线性结构进行汇总,最后总结出数值知识元的描述规则[28],见如表3所示:

表3　数值知识元的描述规则

数值知识元描述规则
时间 + 主体 + 在/从/以/选取/采集/获取/选自/通过/利用/对 + source + (回收/收集/采集/发放/获取/选取/下载/提供/进行/得/为/有/是/达到/有/共计) + 数值 + 单位 + 指标
时间 + 主体 + (从/在/以/选取) + source + (回收/收集/采集/发放/获取/选取/下载/提供/进行/得/为/有/是/达到/有/共计) + 指标 + 数值 + 单位
时间 + 主体 + (最大值/最小值/权重/阈值/维度/临界值/相似值/…率) + (达到/为/非/介于/处于/取/为/大于/等于/小于…) + 数值 + 单位
时间 + 在数值 ~ 数值 + 单位 + 范围内 + 主体 + 指标 + 谓词
时间 + 主体 + (中/过/好/到/有/定/含/内)的((分别/均/仅)(认/设/定/成分/示/本/改/否)为/达到/仅有/下降/上升/提高到/大概为/最低为) + 数值 + 单位
时间 + 数值 + 单位 + 主体 + 指标 + 谓词
时间 + 主体 + (获得/得到/实现/取得) + 数值 + 单位 + 指标

3.1.2　过程型知识元描述规则

(1)方法知识元描述规则。方法知识元的核心主要是介绍方法使用过程、方法使用步骤以及方法使用条件等。关于描述方法使用过程的特征词,其最为明显的则是"首先""然后"等序列性的词。本文针对训练语料中关于方法知识元的描述,大致总结其描述规则如表4所示:

表4　方法知识元描述规则

方法知识元描述规则
…应借鉴/加大/把握/申请/提供/扩大/建立/扭转/强化/细化/加强/及时听取/完善/减少/避免/采用/借助/制定/限定… + 方式/手段/策略/力度
由于/因为…需要/亟待 + 从…,来/以(实现)…
应着手 + 出台/发布/颁布/采取…公告/行动…,来(明确)…
要实现… + 除了要…还要…
(其实现/为了/有效措施/其有效方法)…通过/利用/防止/鼓励/组织…
第一(类/个/中/步/轮/年/阶段/区域/方面)/其一/一要/一是/首先…,
第二(类/个/中/步/轮/年/阶段/区域/方面)/其二/二要/二是/其次…,
第三(类/个/中/步/轮/年/阶段/区域/方面)/其三/三要/三是/最后…
从/在…方面/层面/角度,开展/进行…研究/探讨/调查/讨论/分析

(2)关系知识元描述规则。关系知识元是指所叙

王忠义，沈雪莹，黄京. 基于知识元的中文文本层级分割[J]. 图书情报工作,2019,63(7):105－115.

述对象之间的类属,可以从空间层次和时间逻辑角度将对象之间的关系划分为:并列关系、上下位类属关系、改进关系、演进关系、递进关系、继承关系、替代关系以及因果关系[29]。从静态关系来看,包括并列和上下位类属关系。从动态关系来看,包括改进、演进、递进、继承和替代等关系,这些关系主要特征词为“提出了”“改进”等。因此,关系知识元的描述规则如表5所示:

表5　关系知识元的描述规则

关系知识元描述规则
…＋即/既是/是/并非是…又是/也是/相当于/同时也是/而不是/而是…
…＋是/分为…,一方面/一类是…另一方面/另一类是…
…＋主要包括/包含/有…
…＋可视为/是/是属于…的一部分/组成部分/基础/一支/一种/前提/保障/支柱
…＋从某方面分析,其中…
…(Term/方法)＋如 $M_1,M_2\ M_n$…
…统称为/并列为…的几＋类/大＋…
(根据/按照)…划分为/有…,如/分别是/具体有…
除了＋ Term 还有 Term
基于＋Term,对＋ Term＋提出了/进行了...改进/修正
将/从＋Term＋引入/借用/参考到＋Term＋以/来..
Term＋是＋ Term＋重要原因/因素/要素/动机
对＋Term＋.发挥/造成/产生＋ Term＋影响/推动/作用
Term＋引起/主导/导致/影响/作用于＋Term
一种/一方面/一部分/一类…;另一种/另一方面/另一部分/另一类…

3.2　知识元识别流程

根据上述总结的各类知识元的描述规则,依次与文本中的句子进行匹配,匹配成功,则标记该句子或句群为知识元,否则为起连接作用的衔接句。在对实体资源中的知识元进行识别前,如果待识别的实体资源不是以文本的形式表示,则应该对其进行文档解析,转化为纯文本再进行处理。然后对纯文本进行分词预处理,包括文本分词、句子切分等,并将处理后的文本中的句子以及描述规则按顺序分别存入相应的数据库中,最后利用算法进行匹配,进而识别出知识元。其算法匹配的具体流程如下:

第一步,判断文本库中是否还有其他句子,如果有,则按在文本中的位置顺序选取一个句子,如果没有,转入第五步;

第二步,判断规则库中是否还有其他描述规则,如果有,按顺序选取一个规则,如果没有,则将该句标记为衔接句,并转入第一步;

第三步,将第一步的句子与第二步的描述规则进行匹配,如果匹配成功,则标记为候选句,转入第四步,如果匹配失败,说明该句不符合这一条规则,转入第二步;

第四步,由于匹配过程中,一般是以句子为单位进行匹配的,一个完整句子结束的标志是由“。”“!”“?”等来表示。但是由于某些作者的写作习惯,或者部分句子的特殊性,不能仅仅将一句话作为一个知识元。例如“投入产出分析,又称投入产出核算或部门联系平衡法。它作为一种经济分析方法,从宏观经济角度出发,把国民经济划分成若干不同但互有联系的产品群,借助线性方程,来模拟国民经济结构和社会生产过程,以此综合分析各部门之间的经济技术联系和重要的比例关系”,这两句话实际描述的是一个知识元。所以,为了准确识别文本中描述知识元的范围,在匹配每一个句子时,需要判断候选句后是否有“然后、其、它、这、比如、而且”等一些特殊的连接词、代词、转折词、符号和序列词,若出现,则句子位置后移,将后面的一句加入该句子,视为一个知识元,如果没有出现,则直接标记为该句为知识元,并转入第一步;

第五步,将匹配成功的知识元和衔接句根据它们在文本中出现的位置信息,依次存入数据库中,算法结束。

知识元识别流程图如图3所示:

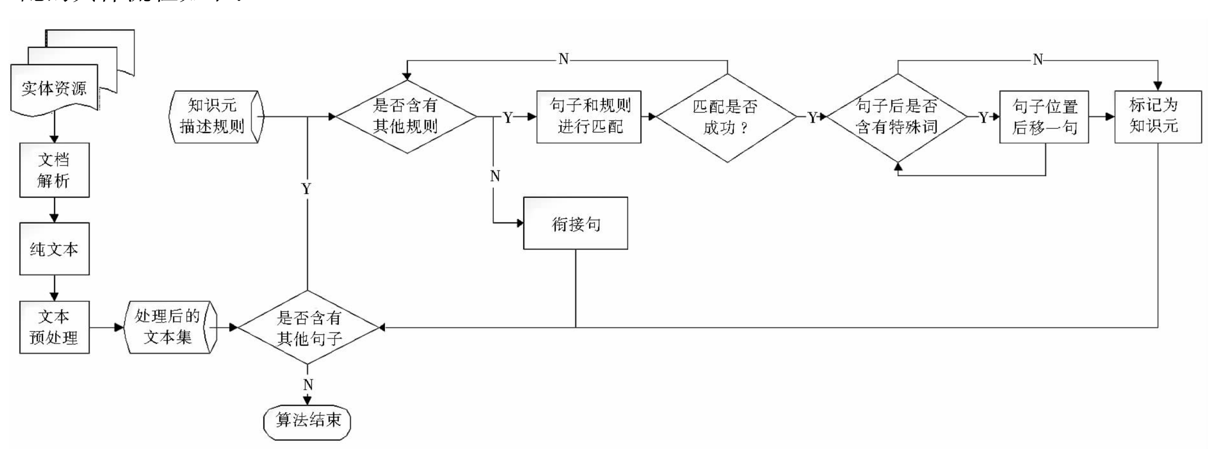


图3　知识元识别流程

4 基于知识元的文本层级分割方法

为便于叙述，本文将知识元和衔接句统称为短文本。基于知识元的文本层级分割方法如图 4 所示。首先基于最长公共子串计算短文本之间的相似度，进而构建短文本相似度矩阵，矩阵中每一行都可以是作为一个短文本向量；然后将待分割的文本中所有的短文本视为一个类，采用 fisher 最优分割法对该类进行逐级二分，直到识别出所有的主题为止，将主题相关的短文本归于一个语义段落，使得语义段落内部具有最大相似性，相邻语义段落之间具有最大相异性，进而识别分割边界，实现文本层级分割。接下来，本文将详细论述各部分的实现过程。

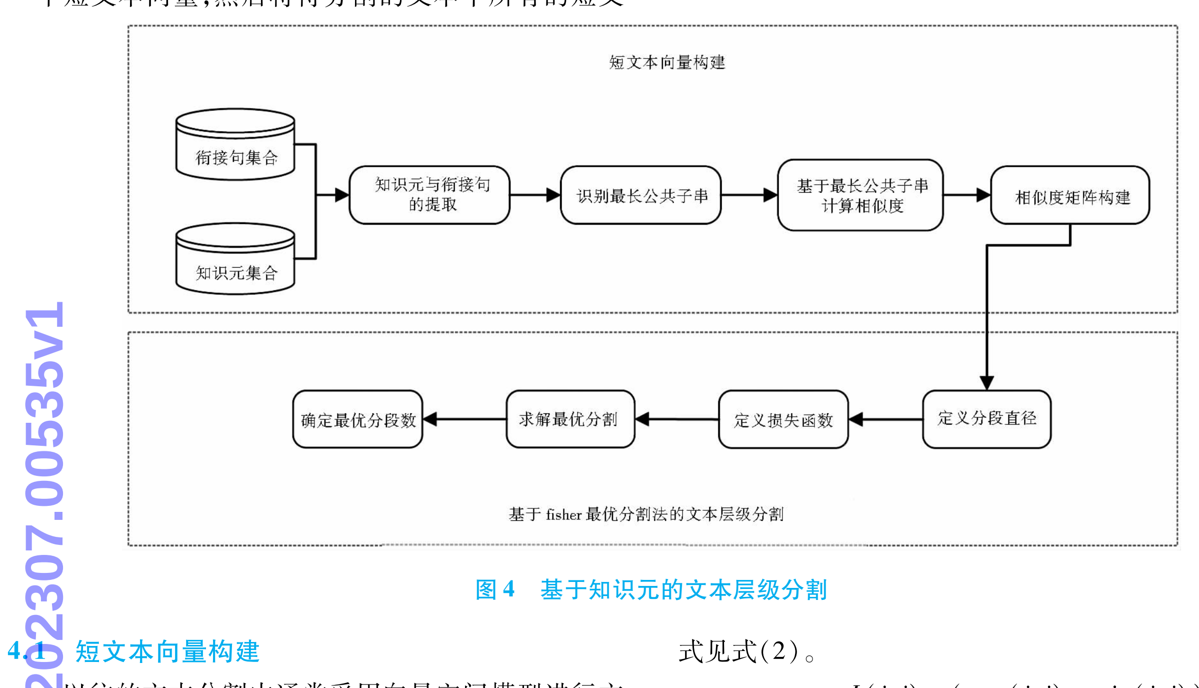


图 4 基于知识元的文本层级分割

4.1 短文本向量构建

以往的文本分割中通常采用向量空间模型进行文本相似度计算，然而，一方面，向量空间模型比较适合于长文本的比较，对于短文本来说，该模型就存在严重的数据稀疏问题；另一方面向量空间模型以词表示文本，未能揭示出词之间的依赖关系，因此具有一定的局限性。为解决这一问题，本文通过计算最长公共子串的方式计算短文本之间的相似度，来构建相似度矩阵。用集合 $G=\{s_1,s_2,\cdots,s_m\}$ 表示含有 m 个短文本的文档，其中 s_i 表示第 i 个短文本，用集合 $s=\{w_1,w_2,\cdots,w_n\}$ 表示包含 n 个词语的短文本，其中 w_i 表示在短文本 s 中第 i 个词语。具体实现过程如下。

首先，设长度分别为 i,j 的两个短文本 $s_1=\{w_{11},w_{12},\cdots,w_{1i}\}$，$s_2=\{w_{21},w_{22},\cdots,w_{2j}\}$ 采用动态规划算法识别两者之间的最长公共子串，其算法见式(1)。

$$L[i,j]=\begin{cases}0 & i=0\ or\ j=0\\ L[i-1,j-1]+1 & w_{1i}=w_{2j}\\ \max\{L[i-1,j],\ L[i,j-1]\} & w_{1i}\neq w_{2j}\end{cases}\qquad 式(1)$$

其中，L[i,j] 表示长度为 i,j 的短文本 s_1,s_2 之间的最长公共子串，w_{1i} 表示短文本 s_1 中的第 i 个词语。

然后，对短文本进行相似度计算，其相似度计算公式见式(2)。

$$sim(s_1,s_2)=\frac{L(i,j)+(\max(i,j)-\min(i,j))}{\max(i,j)}\qquad 式(2)$$

其中，max(i,j) 表示取 i,j 最大数，min(i,j) 表示取 i,j 最小数。

接着，根据短文本之间的相似度来构建相似度矩阵 A。最后，依据矩阵 A，从中提取出各短文本向量。

4.2 基于 fisher 最优分割法的文本层级分割

在对文本进行分割时，为了保证短文本在文本中的写作顺序，本文采用基于 fisher 最优分割法来进行有序分割[30]。具体流程如下。

4.2.1 定义分段直径 令 n 为文本中短文本的总数，即 $s_1,s_2,\cdots,s_n$ 为文档中包含的短文本。因为在文本分割过程中要保留短文本之间的线性顺序，所以分割后的每个类可由 $\{s_i,s_{i+1},s_{i+2},\cdots,s_{i+k}\}$ 表示，记为 $\{i,i+1,\cdots,i+k\}$。设分割后的某一个类为 $\{s_i,s_{i+1},\cdots,s_j\}$，其中 $1\leqslant i<j\leqslant n$，其向量均值为 $\bar{s}_{ij}$，其计算公式见式(3)。类直径为 $D(i,j)$，其计算公式见式(4)。

$$\bar{s}_{ij}=\frac{1}{j-i+1}\sum_{r=i}^{j}s_r\qquad 式(3)$$

$$D(i,j)=\sum_{r=i}^{j}(s_r-\bar{s}_{ij})^T(s_r-\bar{s}_{ij})\qquad 式(4)$$

4.2.2 定义损失函数 将 n 个有序短文本分割为 k

段，设某一种分割后的结果如下：

$\{\{i_1, i_1+1, \cdots, i_2-1\}, \{i_2, i_2+1, \cdots, i_3-1\}, \cdots, \{i_k, i_k+1, \cdots, n\}\}$，其中 $1=i_1<i_2<\cdots<i_k<n$。定义上述分割后的损失函数 L[b(n,k)]为公式(5)。

$$L[b(n,k)] = \sum_{r=1}^{k} D(i_r, i_{r+1}-1) \qquad 式(5)$$

其中，i_1 表示分割 1 中的第一个短文本，i_1+1 表示分割 1 中第二个短文本。当 n,k 固定时，式(5)的值越小，即每个分割段的离差平方和越小，其分割也就越合理。因此，文本层级分隔的目标就转化为寻找一种使得损失函数最小的分割分法 b(n,k)，将其最优分割法记为 p(n,k)。

4.2.3　求解最优分割　为实现文本层级分隔，本文采取逐级二分的策略，对文本进行逐级二分，其二分公式见式(6)。

$$L[b(n,2)] = \min_{2\leqslant j\leqslant n}\{D(1,j-1)+D(j,n)\} \qquad 式(6)$$

选取 j 作为分割点，使得两个分割的离差平方和越小，即分段直径越小，分割也就越合理，然后对新生的两个分割(设为分割 G_1,G_2)分别继续进行二分，如果对 G_1 分割后损失函数大于对 G_2 分割后出的损失函数，则 G_1 不变，选取对 G_2 分割中的分割点，以此类推，继续对生成的所有的类分别进行二分，最终得出最优解 p(n,k)。

4.2.4　确定最优分段数　对文本进行分割时，并不是说将其分的越细越好，需要确定分割段数的阈值 k。然而，在对不同文本进行分割时，不能事先确定文本中包含多少主题，应该生成多少分割。因此，本文通过绘制最小误差函数 L[p(n,k)]随分割段数 k 的变化趋势图 L[p(n,k)]——k(k>1)，在变化趋势图中找出拐点，来作为确定 k 值的依据。需要指出的是，我们不能直接确定拐点相对应的 k 值就是最佳分割段数，仅能将 k 值作为可能的分割段数，即候选分段数。在寻找拐点时，本文通过计算该曲线的斜率差来确定。关于该曲线的斜率差的计算公式见式(7)。

$$\alpha(k) = \left|\frac{L[p(n,k-1)]-L[(p(n,k)]}{k-1-k}\right| - \left|\frac{L[p(n,k)]-L[p(n,k+1)]}{k-(k+1)}\right| \qquad 式(7)$$

当 $\alpha(k)$ 的绝对值达到最大值时，k 值所对应的点即为 $L[p(n,k)]$ 的拐弯处，选取该点及其附近的几个 k 值作为候选分段数。

5　实证

5.1　测试语料选取

根据本文所述的基于知识元的中文文本层级分割方法，笔者从 CNKI 中选取期刊论文《文本分割综述》作为实验测试对象，根据上述的描述规则识别论文中的知识元，然后以知识元为最小单元进行文本层级分割。选取该论文主要因为其包含事实、方法、结论以及数值等多种类型的知识，不同段落之间有着一定的独立性与联系。确定测试对象后，接下来本文采用人工判断的方式来生成文本分割点作为分割标准。笔者邀请 5 位学术研究者来对测试语料进行分割，最后的分割结果遵循少数服从多数的原则，得出标准分割段数为 k=17。

5.2　实验内容

首先，构建数据库，利用上述识别知识元的方法，对测试语料中的知识元进行识别，按其文本顺序将知识元和衔接句存入数据库中。本文所构建的短文本数据库见图 5 所示：

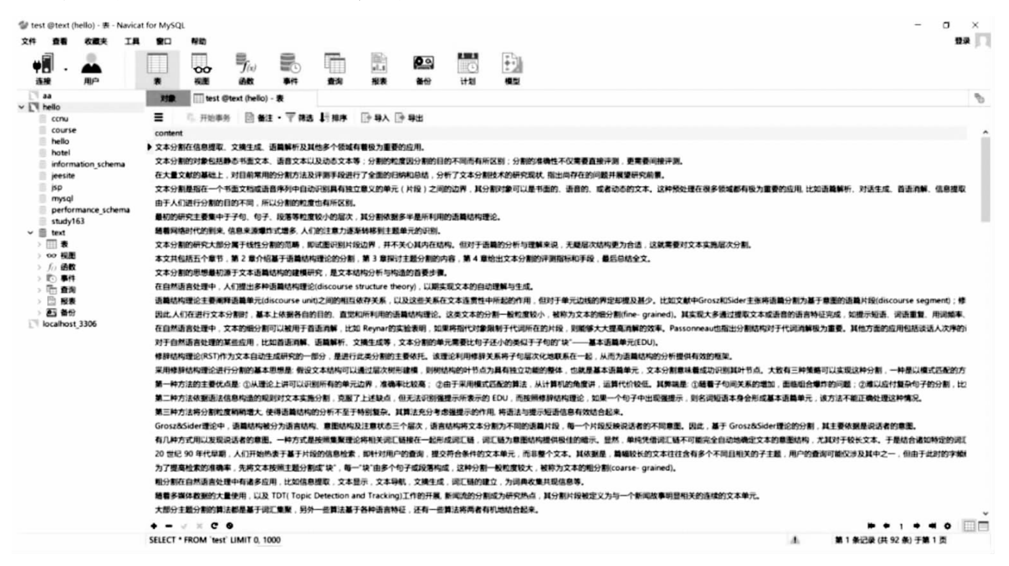


图 5　短文本数据库图

然后，对短文本进行分词预处理，并逐级二分，每次二分后对应的损失函数趋势图见图 6。相应的斜率差变化见图 7。

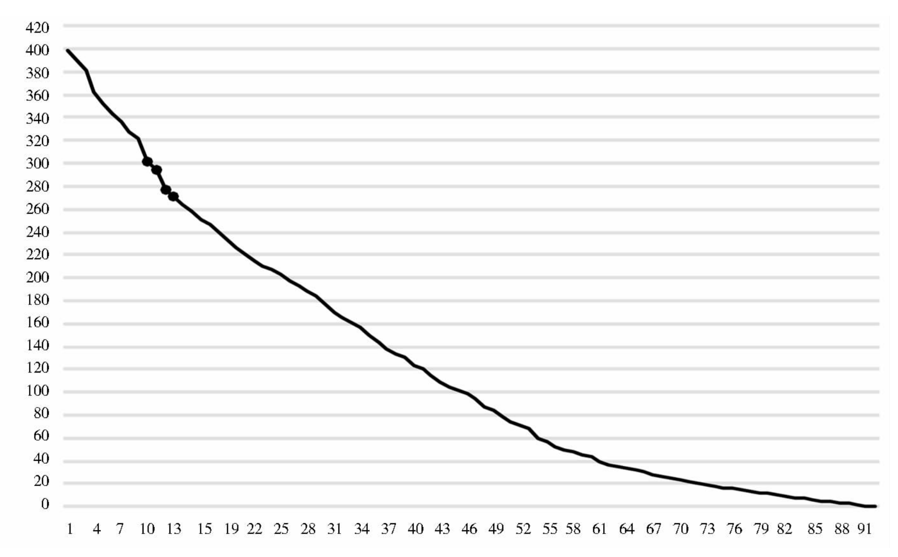


图 6　损失函数变化图

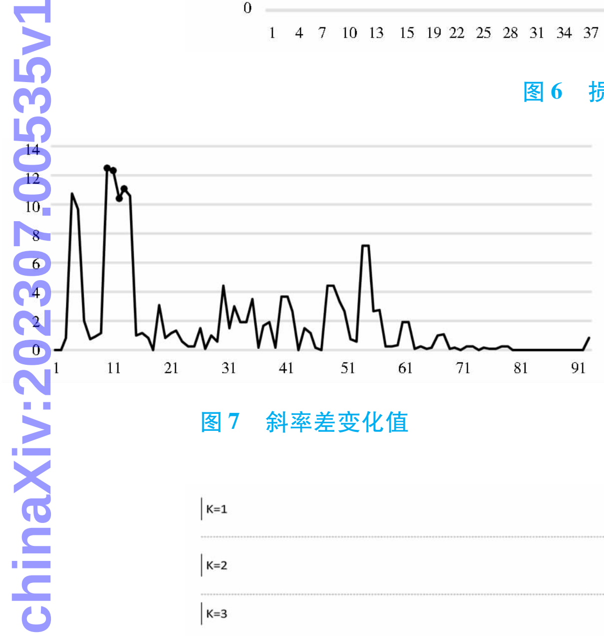


图 7　斜率差变化值

从图 7 中可以看出，斜率差最大时相对应的 k 值为 10，其附近点的斜率差也均大于其他点的斜率差，因此，选取 10，11，12，13 为候选分割段数。从图 6 中可以看出，当 k = 13 时，其相对应的损失函数是最小的，即选取 13 为算法返回的分割段数。其层级分割后的效果图见图 8，其中每一个数字代表一个短文本，该实验对象最后识别出 92 个短文本，阴影部分为最终分割段中所包含的短文本。

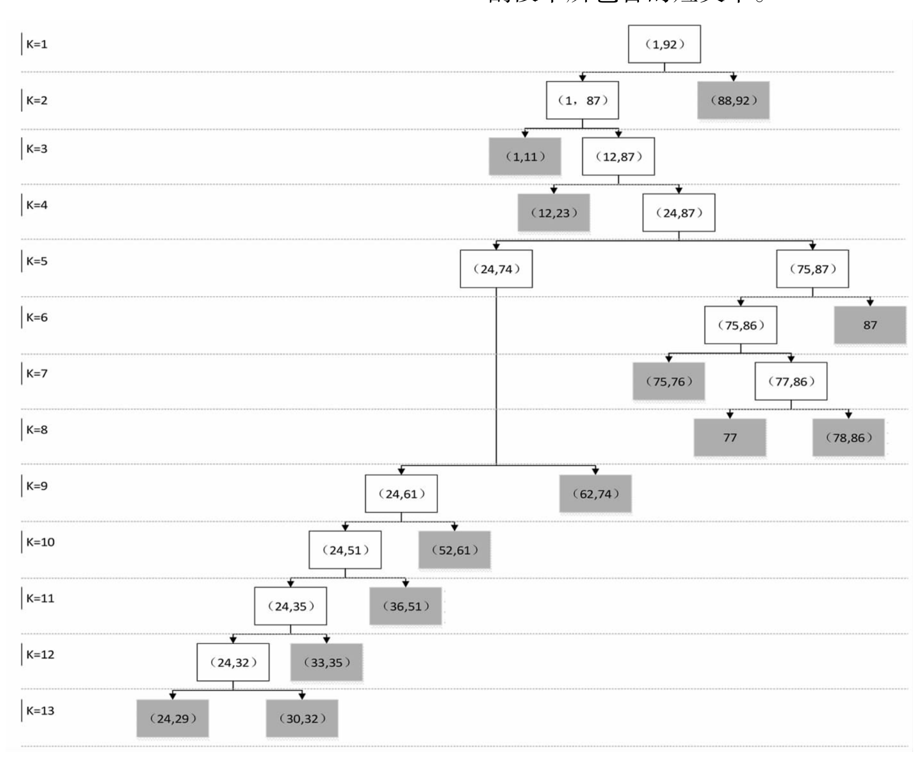


图 8　基于知识元进行文本层级分割效果图

王忠义，沈雪莹，黄京. 基于知识元的中文文本层级分割[J]. 图书情报工作,2019,63(7):105－115.

5.3 测试评价

首先采用传统的正确率 P、召回率 R 以及 F 值来评价该文本分割方法的性能,其计算公式分别见公式(8)、公式(9)和公式(10)。

$$R = \frac{正确识别的分割点数}{文本正确的分割点数} \quad 式(8)$$

$$P = \frac{正确识别的分割点数}{算法返回的分割点数} \quad 式(9)$$

$$F = \frac{2 \times P \times R}{P + R} \quad 式(10)$$

然而,以上这些评估方法只能评价其绝对匹配的结果,而算法返回的边界可能与人工判断的边界只相差一句话。因此,需要进行接近性评价,即本文采用 WindowDiff 评价标准来评价分割结果[31],其计算公式见公式(11)所示。

$$WindowDiff(ref,hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0) \quad (11)$$

其中,在式(11)中,b(i,j)表示短文本 s_i,s_j 两者之间的边界数量,N 表示文本中短文本的总数,ref 代表的是人工分割,认为是标准分割,hyp 代表算法分割,k 表示标准分割中所有分割长度的平均值的一半,WindowDiff 值越低,表明分割算法越准确。

5.4 结果分析

本文选取文献[7]中 M. A. Hearst 提出的经典分割算法 TextTiling 进行实验结果对比,对比结果如表 6 所示。

表 6 对比试验结果

	R	P	F	WindowDiff
本文算法	0.50	0.57	0.58	0.422
TT	0.40	0.50	0.44	0.483

从上表中可以看出,本文算法的准确率和召回率以及 F 均大于 TT 的值(也即:0.5 >0.4,0.57 >0.50,0.58 >0.44),在接近性评价中,本文算法的 WindowDiff 值小于 TT 的值(0.422 <0.483),综合实验数据表明,本文算法是较为合理有效的。其主要原因在于:

首先,本文算法在实现分割时,计算短文本之间的相似度采用的是最长公共子串,该方法有效克服了向量空间模型的数据稀疏与无法描述词语之间依赖关系等问题,其计算相似性更加精准、可靠;其次,本文算法是基于知识元实现的,知识元作为具有相对独立意义的不可再分的最小知识单元,以其为基元进行文本分割可以在逻辑上保证每个分割是一个完整的知识单元,以知识元为单位识别文档分割点时误差也会较小;接着,由于 TT 是通过确定相邻文本块之间的相似性变化程度来确定主题边界,可以实现局部最优,比较适合篇幅较短的文本,对于段落之间差异较大的长文本,TT 则无法通篇考虑其他段落的信息,很难进行正确的判断,而本文算法主要采取的是自上而下逐级二分的策略,在文本分割时可以实现全局最优。因此,基于知识元的文本层级分割的准确率、召回率以及 F 值相对较高,说明本文算法较为合理。此外,在进行接近性评价时,由于本文算法主要是以知识元为单位,进行层级聚类而实现文本分割的,其错误分割一般只在衔接句之间发生,因此,距离正确分割点一般较近,而 TT 算法是以句子为单位进行分割的,其错误分割点可以在任何语句之间发生,有的错误分割点离正确分割点较远,因此 TT 算法的分割误差相对本文算法较大,也说明本文算法相较而言较为合理有效。

6 结语

为了实现对具有层次结构的文档进行层级分割的目的,并提高分割准确性以及效率,本文将文本分割算法的处理单位定为知识元,首先归纳知识元类型以及相关的描述规则,然后根据描述规则中的线索词来识别文本中的知识元,将待分割的文本所包含的知识元和衔接句视为一个类,基于 Fisher 最优算法,对这个类进行逐级二分,直到识别出所有主题,进而形成文本层级结构。该算法有利于将知识服务的控制单位从文献单元深入到知识元、知识元集合为单位,进而来满足用户多粒度的信息需求。通过与经典分割算法 TT 的对比实验,结果表明,本文提出的文本分割算法在精确度指标和接近性评价方面都有着一定的优势的。总体上来讲,利用本文算法对文本进行分割是合理有效和科学的。本文重点关注如何利用知识元来实现中文文本的层级分割,对于知识元描述规则的统计,由于作者写作的习惯以及各个学科描述知识的方法不同,产生的描述规则也不同,因此,本文主要采取人工统计的方法对知识元的描述规则进行梳理统计,尚未探讨自动化抽取过程。

参考文献:

[1] 石晶. 文本分割综述[J]. 计算机工程与应用, 2006, 42(35): 155－159.

[2] REYNAR J C. Topic segmentation: algorithms and applications

[D]. Computer and information science, Philadelphia: University of Pennsylvania,1998.

[3] 邹箭,钟茂生,孟荔. 中文文本分割模式获取及其优化方法[J]. 南昌大学学报(理科版),2011,35(6):597-601.

[4] HALLIDAY M A K, HASAN R. Cohesion in English[M]. London: Routledge,1976.

[5] KOZIMA H. Text segmentation based on similarity between words [C]//Proceedings of the 31st annual meeting on association for computational linguistics. Stroudsburg, PA, USA: association for computational linguistics, 1993: 286-288 .

[6] REYNAR J C. An automatic method of finding topic boundaries [C]//Proceedings of the 32nd annual meeting on association for computational linguistics. Stroudsburg, PA, USA: association for computational linguistics, 1994: 331-333.

[7] HEARST M A. Multi-paragraph segmentation of expository text [C]//Proceedings of the 32nd annual meeting on association for computational linguistics. Stroudsburg, PA, USA: association for computational linguistics, 1994: 9-16.

[8] CHOI F Y Y. Advances in domain independent linear text segmentation[C]//Proceedings of the 1st North American chapter of the association for computational linguistics conference. Stroudsburg, PA, USA: association for computational linguistics, 2000: 26-33.

[9] PONTE J M, CROFT W B. Text segmentation by topic[M]//Research and advanced technology for digital libraries. Heidelberg: Springer Berlin Heidelberg, 1997: 113-125.

[10] MORRIS J, HIRST G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text[J]. Computational linguistics, 1991, 17(1):21-48.

[11] CHOI F Y Y, WIEMER-HASTINGS P, MOOREJ. Latent semantic analysis for text segmentation[J]. Proceedings of emnlp, 2001, 4(3):109-117.

[12] 石晶,戴国忠. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展,2007,44(2):242-248.

[13] 石晶,胡明,石鑫,等. 基于 LDA 模型的文本分割[J]. 计算机学报,2008,31(10):1865-1873.

[14] RIEDL M, BIEMANN C. How text segmentation algorithms gain from topic models[C]//Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg, PA, USA: association for computational linguistics, 2012: 553-557.

[15] EISENSTEIN J, BARZILAY R. Bayesian unsupervised topic segmentation[C]// Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA, USA: association for computational linguistics, 2008: 334-343.

[16] MULBREGT P, CARP I, GILLICK L, et al. Text segmentation and topic tracking on broadcast news via a hidden markov model approach[C]//Fifth international conference on spoken language processing, Sydney, Australia:ISCA Archive,1998:2519-2522.

[17] BRANTS T, CHEN F, TSOCHANTARIDIS I. Topic-based document segmentation with probabilistic latent semantic analysis[C]// Proceedings of the eleventh international conference on information and knowledge management. New York,NY,USA: ACM, 2002: 211-218.

[18] RIEDL M, BIEMANN C. TopicTiling: a text segmentation algorithm based on LDA[C]// ACL 2012 student research workshop. USA: association for computational linguistics, 2012:37-42.

[19] KAN M Y. Combining visual layout and lexical cohesion features for text segmentation[J]. In proceedings of the 31stWorkshop on graph theoretic concepts in computer science- WG 2005, 2001:187-198.

[20] YAARI Y. Segmentation of expository texts by hierarchical agglomerative clustering[EB/OL]. [2018-03-21] https://arxiv. org/pdf/cmp-lg/9709015v1. pdf.

[21] EISENSTEIN J. Hierarchical text segmentation from multi-scale lexical cohesion[C]// Human language technologies: the conference of the North American chapter of the association for computational linguistics. Stroudsburg,PA,USA: association for computational linguistics, 2009: 353-361.

[22] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical dirichlet processes[J]. Journal of the American statistical association, 2006, 101(476):1566-1581.

[23] 李天彩,王波,席耀一,等. 基于分层狄利克雷过程模型的文本分割[J]. 数据采集与处理,2017,32(2):408-416.

[24] 温有奎. 基于"知识元"的知识组织与检索[J]. 计算机工程与应用,2005,41(1):55-57.

[25] 张静,刘延申,卫金磊. 论中小学多媒体知识元库的建设[J]. 现代教育技术,2005,15(5):68-71.

[26] 原小玲. 基于知识元的知识标引[J]. 图书馆学研究,2007(6):45-47.

[27] 赵蓉英,张心源. 基于知识元抽取的中文智库成果描述规则研究[J]. 图书与情报,2017(1):119-127.

[28] 温有奎,温浩,徐端颐,等. 基于知识元的文本知识标引[J]. 情报学报,2006,25(3):282-288.

[29] 化柏林. 学术论文中方法知识元的类型与描述规则研究[J]. 中国图书馆学报,2016,42(1):30-40.

[30] 肖聪,顾圣平,崔巍,等. Fisher 最优分割法在李仙江流域汛期分期中的应用[J]. 水电能源科学,2014(3):70-74.

[31] PEVZNER L, HEARST M A. A critique and improvement of an evaluation metric for text segmentation[J]. Computational linguistics, 2002, 28(1): 19-36.

作者贡献说明:

王忠义:提出研究思路,设计研究方案;

沈雪莹:论文撰写与修改,数据采集,进行实验;

黄京:论文修改与完善。

王忠义, 沈雪莹, 黄京. 基于知识元的中文文本层级分割[J]. 图书情报工作,2019,63(7):105－115.

Chinese Text Hierarchical Segmentation Based on Knowledge Element

Wang Zhongyi[1] Shen Xueying[1] Huang Jing[2]

[1] School of Information Management, Central China Normal University, Wuhan 430079

[2] Wuhan Polytechnic, Wuhan 430074

Abstract: [**Purpose/significance**] This paper aims to help users to retrieve complete and appropriate size of knowledge unit and to satisfy users' multi-granularity requirements. [**Method/process**] This paper proposes a hierarchical segmentation based on the knowledge element. Firstly, the method analyzes the types of knowledge elements and the description rules. Secondly, it identifies the knowledge elements in the entity resources according to the knowledge element description rules, and treats the knowledge elements and the joint sentences as a class. Finally, the fisher segmentation algorithm is used to divide the class bi-levelly until all topics are identified, and the segmentation boundaries are determined, to achieve the hierarchical segmentation. [**Result/conclusion**] This method is based on the recognition of the knowledge element to segment the text. On the one hand, segmentation granularity extends from sentence to knowledge element, which improves the efficiency of segmentation. On the other hand, the control unit of knowledge service is deepened from the literature into knowledge blocks with knowledge elements and knowledge elements sets as the unit, providing the necessary knowledge resources, realizing the progress from data retrieval, information retrieval to knowledge retrieval, improving the efficiency of knowledge acquisition and achieving the transformation of information services to knowledge services.

Keywords: knowledge-element recognition clustering hierarchical segmentation